## Chi-Square Test

→ Chi-Square test is a statistical test used to determine whether there is a significant association b/w categorical variables. It is particularly useful for analyzing data where you have two or more categorical variables and you want to assess whether they are independent or related in some way. There are several variation of chi-square test.

i) Chi-Square ~~Table~~ Test for Independence :- This test is used to determine if there is significant association b/w two categorical variables in contingency table. It helps answer question like "is there a relationship b/w gender and voting preference"

ii) Chi-square Goodness of fit test :- This test assesses whether an observed frequency distribution of categorical data matches an expected theoretical distribution. For eg, you might use it to check if the distribution of blood types follows expected proportions.

iii) Chi-square test for Homogenity :- This test is used to compare the distributions of categorical variables across different groups or populations to see if they are significantly different.

# CHI- SQUARE TEST FOR INDEPENDENCE

Q. Suppose you are conducting a survey to investigate whether there is an association b/w a person's level of education & their preferred type of vacation destination. You collect data from 300 individuals and categorize them into 3 education levels ("High School", "Bachelor's Degree", and "Masters Degree") and three vacation preferences ("Beach", "City", "Mountain"). Your goal is to determine if there is a significant association b/w education level and vacation preference.

| | Vacation Preference | | |
|---|---|---|---|
| Education level | Beach | City | Mountain |
| High School | 50 | 30 | 20 |
| Bachelor's Degree | 40 | 60 | 10 |
| Masters Degree | 30 | 40 | 30 |

i) Is there a significant association b/w a person's level of education & their preferred type of vacation destination? Conduct a Chi-square test for Independence at a 5% significance level.

## ANSWER

Step 1: Set up Hypothesis

- Null Hypothesis ($H_0$): There is no association b/w education level & vacation preference.

  $H_0$:- Two categorical variables are independent.

- Alternative Hypothesis ($H_1$): There is an association b/w education level and vacation preference.

  $H_1$:- The two categorical variables are not independent.

Step 2 :- Choosing Significan level.

$\alpha = 0.05$.

Step 3 :- Create Contigency table:
(Look at the table)

Step 4 :- Calculate Expected frequencies

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

| Education level | Beach | City | Mountain | Total |
|---|---|---|---|---|
| High School | 50 | 30 | 20 | ⇒ 100 |
| Bachelor's Degree | 40 | 60 | 10 | ⇒ 110 |
| Masters Degree | 30 | 40 | 30 | ⇒ 100 |
| | 120 | 130 | 60 | 310 |

↓ Grand total

$\boxed{\text{Expected frequencies}}$

| Education level | Beach | City | Mountain |
|---|---|---|---|
| High School | 38.71 | 41.93 | 19.35 |
| Bachelor's Degree | 42.58 | 46.129 | 21.29 |
| Master's Degree | 38.7 | 41.935 | 19.35 |

Step 5 :- Calculate Chi-square Statistic

$$\chi^2 = \Sigma \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \frac{(11.29)^2}{38.71} + \frac{(1.93)^2}{41.93} + \frac{(0.65)^2}{19.35} + \frac{(2.58)^2}{42.58} + \frac{(13.87)^2}{46.129}$$

$$+ \frac{(11.29)^2}{21.29} + \frac{(8.7)^2}{38.7} + \frac{(1.935)^2}{41.935} + \frac{(10.65)^2}{19.35} =$$

$= 3.29 + 3.394 + 0.0218 + 0.1563 + 4.17 + 5.987 + 1.955 +$
$0.089 + 5.86$

$= \underline{24.923}$

Step 6 :- Determine Degree of freedom.

$df = (\text{No. of Rows} - 1) \times (\text{Number of columns} - 1)$

$= \underline{4}$

Step 7. Calculate Critical value

for, $d_f = 4$ & $\alpha = 0.05$ look at Chi-Square table
critical value $= 9.488$

Step 8 :- Compare critical value & Chi-Square Statistic

If $\chi^2 \leq$ Critical Value, we fail to reject null hypothesis
$\chi^2 >$ Critical value, reject null hypothesis.

Here $\chi^2 >$ Critical value
$\therefore$ we reject Null hypothesis.

Conclusion: At 95% CI & 5% Significance level we can say that the 2 categorical variables are not independent. There is a significant association b/w person's level of education & vacation destination.

# CHI - SQUARE GOODNESS of FIT TEST

Q. Imagine you are a Quality control Manager at a candy manufacturing company. Your company produces bags of assorted candies, and you claim that the bags contain an equal distribution of four different flavors: "Chocolate", "Strawberry", "Lemon", "Mint". However a consumer group suspects that your bags are not equally distributed and that one of the flavors is underrepresented. To investigate, you randomly select 200 candies from a bag and record the number of each flavor you find. You want to test if the observed distribution of flavors match your claim.

Observed distribution
Chocolate - 60
Strawberry - 50
Lemon - 45
Mint - 45

Question
Is there a significant difference b/w the observed distribution of candy flavors and the claimed distribution (equal distribution of 25% for each flavor)? Perform a Chi-square Goodness of fit Test at 5% Significance level.

$\boxed{\text{ANSWER}}$

Step 1 :- Set up Hypothesis
$H_0$ :- The observed distribution matches the claimed distribution (equal distribution of 25% of each flavor)
$H_0$ :- The distribution of flavor is as claimed
$H_1$ :- Distribution of flavor is not as claimed

Step 2 :- Expected Distribution.

Chocolate → $0.25 \times 200 = 50$

Strawberry $= 50$

Lemon $= 50$

Mint $= 50$

Step 3 :- Calculate Chi-Square Statistic.

$$\chi^2 = \Sigma \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Chocolate $= \cancel{60}\ \frac{100}{50} = 2$

Strawberry $= 0$

Lemon $= \frac{25}{50} = 0.5$

Min $= 0.5$

$\chi^2 = \frac{3}{2}$

Step 4 :- Degree of freedom $= 3$

Step 5 :- Find critical value from Chi-square table $= 7.815$

Step 6 :- Conclusion

$\chi^2 <$ critical value we fail to reject Null hypothesis

# CHI-SQUARE TEST FOR HOMOGENITY

Scenario :- Suppose you work for a political polling organization and you want to determine whether there is a significant difference in political party preferences among three different age groups: "Young Adults" (ages 18-30), "Middle-Aged Adults" (ages 31-50), and "Senior Adults" (ages 51 & above). You conduct a survey and record the political party preference of 300 individuals in each age group. You want to test if the distribution of political party preferences is the same across all three age groups.

Data:

Here's the summarized data:

- For "Young Adults" (ages 18-30):
  90 individuals prefer Party A
  110 Individuals prefer Party B
  100 Individuals prefer Party C

- For "Middle-Aged Adults" (ages 31-50)
  70 individuals prefer Party A
  120 individuals prefer Party B
  110 individuals prefer Party C

- For "Senior Adults" (ages 51 and above)
  60 individuals prefer Party A
  90 individuals prefer Party B
  150 individuals prefer Party C

- Question:- Is there a significant difference in political party preferences among the three age groups: "Young Adults", "Middle-Aged Adults", and "Senior-Adults"? Perform a Chi-Square Test for Homogenity at 5% significance level.

## ANSWER

Step 1. $H_0$ :- The distribution of political party preference is same across all age groups.

$H_1$ :- Distribution of political party preference is different among age groups.

Step 2. Contingency table.

| Age Group | Party A | Party B | Party C | |
|---|---|---|---|---|
| | Political Party Preference | | | |
| Young Adults | 90 | 110 | 100 | 300 |
| Middle-Aged Adults | 70 | 120 | 110 | 300 |
| Senior Adults | 60 | 90 | 150 | 300 |
| | 220 | 320 | 360 | 900 |

Step 3. Chi - Square Statistic

~~Expected~~ Observed ~~values~~ proportion of Party A = $\frac{90+70+60}{900} = \frac{220}{900} = 24.4\%$

Expected value for Young Adults of Party A = 24.4% of 300

| Age Group | Party A | Party B | Party C |
|---|---|---|---|
| Young Adults | 73.33 | 106.67 | 120 |
| Middle-Aged Adults | 73.33 | 106.67 | 120 |
| Senior Adults | 73.33 | 106.67 | 120 |

↓

$$\text{Expected Value} = \frac{\text{Row (total)} \times \text{Column (total)}}{\text{Grand Total}}$$

$$\chi^2 = 3\times\left(\frac{(90-73.33)^2}{73.33} + \left(\frac{(110-106.67)^2}{106.67} + \frac{(120-100)^2}{120}\right)\right)$$

$$= 3\times(3.789 + 0.1039 + 3.33)$$

$$= 21.668$$

Step 4. Finding critical value for dof = 4 & $\alpha = 0.05$, 9.488

Step 5. Conclusion :- Since $\chi^2$ > Critical value we reject null hypothesis. Distribution of political party preference is diff. for age groups.